Recognizing Emotions in Video Using Multimodal DNN Feature Fusion

Recognizing Emotions in Video Using Multimodal DNN Feature Fusion

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu
Centre for Speech Technology Research (CSTR)
University of Edinburgh, UK
j.williams@ed.ac.uk

Abstract

We present our system description of input-level multimodal fusion of audio, video, and text for recognition of emotions and their intensities for the 2018 First Grand Challenge on Computational Modeling of Human Multimodal Language. Our proposed approach is based on input-level feature fusion with sequence learning from Bidirectional Long-Short Term Memory (BLSTM) deep neural networks (DNNs). We show that our fusion approach outperforms unimodal predictors. Our system performs 6-way simultaneous classification and regression, allowing for overlapping emotion labels in a video segment. This leads to an overall binary accuracy of 90%, overall 4-class accuracy of 89.2% and an overall mean-absolute-error (MAE) of 0.12. Our work shows that an early fusion technique can effectively predict the presence of multi-label emotions as well as their coarse-grained intensities. The presented multimodal approach creates a simple and robust baseline on this new Grand Challenge dataset. Furthermore, we provide a detailed analysis of emotion intensity distributions as output from our DNN, as well as a related discussion concerning the inherent difficulty of this task.

1 Introduction

Automatic emotion detection is a longstanding and challenging problem in the field of artificial intelligence and machine learning. One reason why emotion analysis is so difficult is due to the fact that emotions are somewhat subjective, which affects how emotions are perceived and subsequently labeled by human annotators. To compound this even further, the expressed emotions may change, in particular for video data. In addition, multiple emotions can be expressed simultaneously and also as a sequence over time. Emotions provide a type of para-linguistic information that is crucial for many applications in artificial intelligence including: affective speech generation, bio-medical diagnostics, machine translation and human-computer interaction.

Multimodal machine learning has been recently attracting interest, with the abundance of multimedia data available on the internet making it easy for researchers to integrate data of multiple modalities. It is a dynamic research field which aims to integrate and model multiple sources of input, usually acoustic, visual and text.

In order to produce major advances in emotion analysis, there must be adequate techniques for combining and analyzing complex signals. While this notion is applicable across many fields and tasks, in this work we focus on emotion analysis from video data — a very active research area that is beaming with interesting results and methodologies (Pérez-Rosas et al., 2013; Wöllmer et al., 2013; Poria et al., 2015; Brady et al., 2016; Zadeh et al., 2016b). A survey by Baltrušaitis et al. (2018) motivates some of the uses of multimodal analysis, together with five main components:

- **Representation:** Representing and summarizing multimodal data
- **Translation:** Mapping data from one modality to another
- **Alignment:** Identifying relationships between modalities: for example, transcribed text of a video
- **Fusion:** Joining information for different modalities in order to perform a prediction
- **Co-learning:** Exchanging knowledge between modalities

Our work touches on representation, alignment, and co-learning issues, but it is mostly focused on fusion. Specifically, we are interested in finding a way to predict emotions from video data by fusing together three modalities: verbal content, acoustic features and sequences of images. In this work we provide the experimental framework for developing a system for 6-class (multi-label) emotion classification and regression for the First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language at Association for Computational Linguistics (ACL) 2018. [1]

This paper is organized as follows: in Section 2, we present some relevant work on multimodal emotion recognition. In Section 3 we provide an overview of the CMU-MOSEI dataset and a description of our task. In Section 4, we present our methodology and multimodal fusion technique. In Section 5, we show our experiments and results. In Section 6 we show some analysis of our experiments and in Section 7 we finally discuss and make suggestions for future work.

2 Related Work

In light of recent successes with deep learning approaches to multimodal classification problems (Zadeh et al., 2017), emotion analysis remains truly challenging. Both emotion and sentiment analysis have become increasingly important in recent years. However, it remains a difficult task due to the ambiguity of language and the use of slang and sarcasm (Baltrušaitis et al., 2018; Poria et al., 2017; Soleymani et al., 2017). A persistent idea is that information from other modalities helps to resolve ambiguities, such as adding information about facial features. From the first time that convolutional neural networks (CNNs) were employed for face recognition (Lawrence et al., 1997) to the present times when sentiment analysis revolves around using CNNs (Tripathi et al., 2017; Xu et al., 2014; Pereira et al., November 2016), CNNs appear promising for multimodal sentiment analysis and emotion recognition.

One way to encourage innovation in the area of multimodal emotion analysis is through annual shared tasks. One such task is the Audio Video Emotion Challenge (AVEC) which encourages creative and robust approaches to multi-signal emotion recognition. In 2016, the top-performing emotion recognition system utilized sparse coding as well as a state space estimation approach to multimodal fusion (Brady et al., 2016). Similar to our approach, they used both convolutional networks (CNNs) and recurrent neural networks (RNNs). Their system competed internationally and achieved the top scores for valence and arousal. However, their work was slightly different from ours in that they were working with a different set of signal modalities (audio, video and electro-cardiogram (EEG)) and predicting emotion continuously over time. In addition, the AVEC 2016 Challenge relied on a very small pool of subjects. Our work is based on more than 80 different speakers and our prediction task for videos is conducted on a per-segment basis.

Previous work has shown that there are particular elements of the speech signal which are most indicative of emotional state of the speaker (Chang et al., 2011; Zeng et al., 2009). The features of speech which are most predictive of speaker affect are called low-level descriptors. These low-level descriptors can be extracted from the audio signal using a standard speech toolbox such as the COVAREP software (Degottex et al., 2014).

Speech data is often considered sequentially informative. For example, the rise and fall of prosody can form meaningful patterns. Many approaches to detecting emotion in speech use recurrent neural network (RNN) approaches to sequential learning, such as Long-Short Term Memory (LSTM) (Lim et al., 2016). There has been work on emotion recognition using Bidirectional LSTMs, which we also use for developing our best system (e.g. Ghosh et al., 2016; Lee and Tashev, 2015; Han et al., 2014; Chernykh et al., 2017).

There is also considerable work in the area of multi-label emotion recognition for music where the multi-label task has been transformed into sets of one-vs-all (Trohidis et al., 2008). While that approach can be very useful for similar multi-label tasks, we show that our algorithmic approach using DNNs overcomes the need to transform the problem into one-vs-all. Furthermore, we note that there are many ways to evaluate multi-label recognition tasks; in this work however, we followed the metrics set forth by the organizers.

One dataset in particular, called IEMOCAP, is commonly employed for emotion recognition research. It was developed by eliciting specific emotions from subjects while they were being monitored. For example, their facial expressions and

[1] http://multicomp.cs.cmu.edu/acl2018multimodalchallenge/

hand movements were recorded while they spoke. The subjects functioned as emotional actors and were asked to perform scripts that were designed to elicit specific emotions: happy, angry, sad, frustrated and neutral (Busso et al., 2008). However, our work uses a slightly broader set of emotions and multiple emotion labels can be activated simultaneously. More importantly, our data is from speakers who have exhibited emotions spontaneously and, according to their own inclination, similar to real-world contexts.

3 Data and Task

In this section we describe the data that we used for developing our Grand Challenge emotion recognition system and more details related to our prediction task.

3.1 Data Description

In an effort to overcome the challenge of consistent emotion labeling, and to allow for meaningful comparison across systems, our work is based on a standardized emotion dataset, called CMU-MOSEI (Zadeh et al., 2018), from the CMU-MultimodalDataSDK toolbox.[2] This dataset contains video segments that were collected 'in the wild' from YouTube wherein the speaker is providing their review of a movie that they have seen. The segments have been labeled by humans for 6 different emotions, including the null case. These labels are: $Anger$, $Disgust$, $Fear$, $Happy$, Sad, and $Surprise$. Each segment can have any combination of emotion labels, or no labels at all. In addition, for each emotion label there is a corresponding regression value in the range of $[0, 3]$ in 9 steps, making step sizes of approximately 0.33 or $1/3$. This means that every video segment can be characterized with an emotion as well as the intensity of that emotion.

The CMU-MOSEI dataset (Zadeh et al., 2018) provides pre-processed features and a way to align features; we aligned the data to text throughout all experiments. We chose this because the code for this alignment method was already provided by the CMU-MultimodalSDK toolbox.

Text features consist of word vectors obtained from the Global Vectors for Word Representation (GloVe) software (Pennington et al., 2014) as well as one-hot word representations.

[2]https://github.com/A2Zadeh/CMU-MultimodalDataSDK

Audio features were extracted using the software COVAREP: 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients. The sampling rate of these features is 100 Hz from the original audio (Degottex et al., 2014)

Video features were extracted using the Emotient FACET software (Littlewort et al., 2011). According to Zadeh et al. (2016a), the visual features include 16 Facial Action Units, 68 Facial Landmarks, Head Pose and Orientation, 6 Basic Emotions and Eye Gaze (Wood et al., 2015; Baltrusaitis et al., 2014). FACET provides frame-by-frame tracking of facial action units. These features are sampled at 30 Hz.

The most common target emotion in our training data is the singleton $Happy$, followed by the $null$ class and the Sad class. The emotion labels can be combined in various ways. For example, the tuples: $(Happy, Sad)$ and $(Anger, Happy)$ both occur with relatively high frequency and are more frequent than the singleton $Fear$.

3.2 Task Description

Using the CMU-MOSEI dataset, we identified our best-performing early fusion prediction system for the emotion recognition Grand Challenge. While the challenge dataset contains emotion labels as well as sentiment labels, our present work is focused entirely on emotion recognition.

Overall our task was to simultaneously predict **emotion label** (none, one, or many) as well as the corresponding **emotion intensity** for each video segment using a fusion of modalities. The exemplar targets can be visualized as follows:

$target = [0., 0., 0.33, 0.66, 0., 0.]$

where the array indexes correspond to the set of 6 emotion labels and the continuous values (in steps of 0.33) correspond to intensity. In the above example there are two emotions present simultaneously for this video segment $(Happy, Sad)$, and the two emotions differ in their intensity.

First, we created our own custom data split from the CMU-MOSEI challenge data so that we could utilize a held-out test set. This custom split allowed us to train, validate, and test various ablation groups, compare our models, and identify the best-performing system to use for the emotion recognition Grand Challenge. Otherwise our submission for the Grand Challenge would have re-

lied solely on the performance of a validation set, which may have led to unintentional overfitting when comparing several models.

With our custom split, we had the following distribution of examples: Training: 9400, Validation: 1800, and Testing: 1100, for an approximate split of 76/14/10. To this end, we used our custom data split to experiment with unimodal systems, bimodal systems, and trimodal systems, before submitting our final best-performing model to the Grand Challenge. We used overall mean-absolute error (MAE) as a metric for determining the best model. Finally, our actual system submission to the emotion recognition Grand Challenge was trained, validated, and tested on the standardized data split as provided by the organizers.

4 Methodology

In this section we outline our methodology. First, we describe each of the DNNs that we considered, followed by an explanation of how our system design for input-level multimodal fusion (i.e. early fusion) works. Finally, we provide details regarding feature alignment and DNN hyper-parameters.

4.1 DNN Architectures

CNN: Convolutional Neural Networks are often used in NLP for various prediction tasks, including sentiment analysis (Kim, 2014). The interpretation is not as straightforward as for images, but we can still argue that semantically related vectors will be close to each other within a context window. As outlined later in the methodology, we use one-dimensional Convolutional layers.

LSTM: Recurrent Neural Networks (RNNs) and variants have been proven very successful for many tasks including sentiment analysis on text and are known for their ability to model invariances across time. Recent advancements propose variants of RNNs that do not suffer from the problem of vanishing gradients: Long Short Term Memory (LSTM). The goal of LSTMs is to capture long distance dependencies in a sequence, such as the context words.

Bidirectional LSTM: Bidirectional LSTMs (BLSTMs) increase the amount of available contextual information. The principle is to use both a forward pass and a backward pass through, for instance, a video segment, while treating the features as meaningfully sequential.

4.2 Early Fusion

In the early fusion approach, features from each of the 3 modalities are concatenated at the input-level and together they become the input vector to a DNN — this approach is shown in Figure 1. Since sequences have different lengths, all modalities are processed with a maximum cutoff, in order for the concatenation to be possible. We chose the optimal value for the maximum cutoff by exploring a range of values during the hyper-parameter search. The concatenated features are then fed into a DNN.

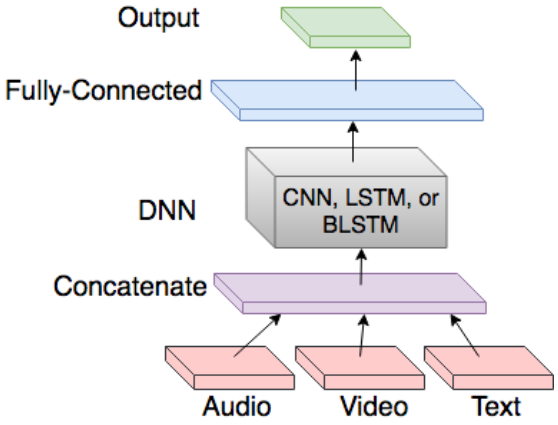

Figure 1: Input-level feature fusion architecture.

4.3 Feature alignment

For our bimodal and trimodal experiments, we align the modalities, because different features in multimodal datasets are in different temporal frequencies. The CMU-MultimodalSDK toolbox aligns data using weighted averaging. The overlap of each modality with a reference one is the weight of each modality. An average is taken with these weights to align them to the reference.

4.4 DNN Hyper-parameters

All of our experiments were trained using the Keras Library (Chollet et al., 2015) which is based on Tensorflow (Abadi et al., 2016). Across all of our experiments, we used the ReLU (Nair and Hinton, 2010) activation function to introduce non-linearity. The learning rule was Adam (Kingma and Ba, 2014) with default Tensorflow parameters. For 1D convolution layers the kernel size was 3 and for max pooling layers the window size was 2. We explored the number of layers in steps of 1, 2 and 3, for both fully connected layers and convolutional layers. For LSTMs and

Bi-directional LSTMs we set the number of units to 64 and for all fully connected layers we set the number of units to 100.

We added dropout (Srivastava et al., 2014) between fully connected layers with dropout rate in $\{0.1, 0.2\}$. We varied the maximum length setting for the video segments in our dataset, known as $maxlen$, in $\{15, 20, 25, 30\}$. We chose these values for maximum length cutoff based on the average segment length reported in Zadeh et al. (2016b), which was indicated as $maxlen = 12$.

In all experiments we used early stopping with the stopping criteria set to identify minimum validation loss and patience was set to 10. The experiments employed batch normalization with batch sizes set to 64 (Ioffe and Szegedy, 2015). The final output layer contained 6 neurons, followed by a linear activation function that bounded values between 0 and 3.

The loss was measured via the mean-absolute error (MAE), where smaller values are better and zero is considered perfect. Our interpretation of MAE is that a value below 0.166 or $1/6$ is considerably good performance, based on the intensity range of $[0, 3]$ and the step size of 0.33. Later, we shall describe additional evaluation metrics that were used with our Grand Challenge submission.

5 Experiments

In this section we present the results of our experiments on a random prediction baseline, followed by unimodal, bimodal and trimodal input-level feature fusion. We used the outcome of these experiments to evaluate and compare each model perfomance. Finally, we provide the results for the Grand Challenge from our best-performing system: the trimodal BLSTM.

5.1 Random Baseline

Developing a baseline was motivated by the fact that this is the first shared-task on the CMU-MOSEI dataset, and therefore no existing systems are available for a direct comparison. There are several different ways of developing a baseline on this task: (1) fully-randomized, (2) preserving label-category distributions from training data or (3) preserving label-quantity distributions from training data. We developed a fully-randomized baseline because it is the most trivial model. Our random baseline methodology can be easily adapted to other metrics used by the shared-task organizers, such as 4-class accuracy.

Emotion	MAE
Anger	0.70
Disgust	0.68
Fear	0.62
Happy	0.80
Sad	0.72
Surprise	0.05
Overall	**0.60**

Table 1: Baseline MAE based on randomized predictions of quantity of labels, label category, and intensity.

First, we generated a random number n for the quantity of labels present in a given video segment from the domain $n = \{0, 1, 2, 3, 4, 5, 6\}$ so that none or all emotion labels could potentially be predicted. Given this quantity, we predicted the identity of the labels by randomly choosing n labels from the domain $[Anger, Disgust, Fear, Happy, Sad, Surprised]$. Finally, we randomly predicted an intensity for each label based on the 9-step regression values in the range of $[0, 3]$, with step size 0.33. The result was an array for each video segment which we used to compare with the truth labels in our small, held-out test set. Table 1 displays our per-label prediction values in terms of MAE. Therefore we can say that if a system performs better than overall MAE of 0.60 (lower values are better) then it is performing better than pure chance.

5.2 Unimodal

To begin with, we experimented with unimodal approaches to set another performance baseline and to find out if any particular modality seemed to contribute significantly more, or if performance was skewed. The results for unimodal performances of each DNN can be found in Table 2. We used our custom training/validation/test split of the available data to obtain this performance, where the overall MAE is only reported on a small held-out test set (but not the official Grand Challenge test set). The performance metric MAE has been averaged over all of the 6 emotion label classes.

The audio modality performed best with a CNN. On the other hand, both text and video performed better with LSTMs. This suggests that text and video provide learnable structures that are captured with sequence modeling.

Modality	DNN	Overall MAE
Audio	LSTM	0.150
	BLSTM	0.150
	CNN	**0.146**
Video	**LSTM**	**0.146**
	BLSTM	0.147
	CNN	0.149
Text	**LSTM**	**0.156**
	BLSTM	0.157
	CNN	0.158

Table 2: Unimodal prediction results, overall mean-absolute error (MAE) for each DNN.

Modality	DNN	Overall MAE
Audio+Video	LSTM	0.137
	BLSTM	**0.135**
	CNN	0.138
Audio+Text	**LSTM**	**0.140**
	BLSTM	0.142
	CNN	0.146
Text+Video	LSTM	0.149
	BLSTM	**0.145**
	CNN	0.149

Table 3: Bimodal prediction results, overall mean-absolute error (MAE) for each DNN and ablation.

5.3 Bimodal

For each bimodal ablation group model, we combined two of the three modalities with a DNN. We report the results in Table 3. We used our custom train/valid/test split of the available data to obtain this performance. We observe that overall, the bimodal ablations performed slightly better than single modalities in terms of overall MAE. The audio+video ablation group performed better than other modality pairs. This could be related to the ambiguity of spoken language. Emotions that embody sarcasm, irony, and typical spoken disfluencies may be better captured without the noise of the text. Text can be particularly misleading in cases of sarcasm, where the truth-value of a sentence is reverse from its literal interpretation.

5.4 Trimodal

We present the results of our trimodal fusion in Table 4. Once again, we used our custom training/validation/test split of the available data to obtain this performance. It is interesting to note that all of these systems performed similarly well, and all performed better than the bimodal ablation groups. Based on the results from our trimodal experiments, we selected the BLSTM to submit as our system to the Grand Challenge.

DNN	Modality	Overall MAE
LSTM	A,V,T	0.133
BLSTM	**A,V,T**	**0.132**
CNN	A,V,T	0.134

Table 4: Trimodal prediction results, overall MAE for each DNN. Note A=Audio, V=Video, and T=Text.

5.5 Grand Challenge Results

To obtain the official Grand Challenge results, we trained our BLSTM using the original dataset split as provided by the organizers for training and validation. We then applied our system model to an unseen test set and submitted our predictions. The evaluation results were returned to us by the challenge organizers.

Our system performance is displayed in Table 5. It shows the performance on a per-emotion basis as well as the overall metric. We noticed that our system's overall performance, in terms of MAE, on this held-out test set was slightly better than what we obtained while constructing our model during earlier experiments. This could be due to the fact that we used the entire provided training and validation set for the submission.

First, binary accuracy was calculated by rounding values to the nearest integer, and using non-zeros for the 'positive' class and zeros as the 'negative' class. Binary accuracy is used to measure the presence and absence of an emotion label. Next, the 4-class accuracy is obtained in a similar way. Each value is rounded to the nearest integer in $\{0, 1, 2, 3\}$ resulting in 4 classes. And the accuracy is again measured on exact matches. The 4-class accuracy provides a rough estimate of how well a system predicts intensity of an emotion because the 4-classes provide a coarser-step size within the range of regression values (e.g. 4 steps in the range [0,3] instead of 9 steps). Finally, the correlation r is provided for a fine-grained metric that measures how well the system output correlates with the true intensities from the data.

For each emotion label, our correlation values are near 0, which indicates that our system outputs do not correlate with fine-grained emotion inten-

Emotion	MAE	Binary Acc(%)	4-class Acc(%)	Corr. r
Anger	0.101	92.6	92.6	0.082
Disgust	0.051	96.3	96.3	0.064
Fear	0.051	95.7	95.7	0.011
Happy	0.404	70.5	62.0	0.551
Sad	0.111	91.0	91.0	-0.062
Surprise	0.038	97.7	97.7	-0.030
Overall	**0.126**	**90.6**	**89.2**	–

Table 5: Official Grand Challenge system results for our early-fusion trimodal BLSTM.

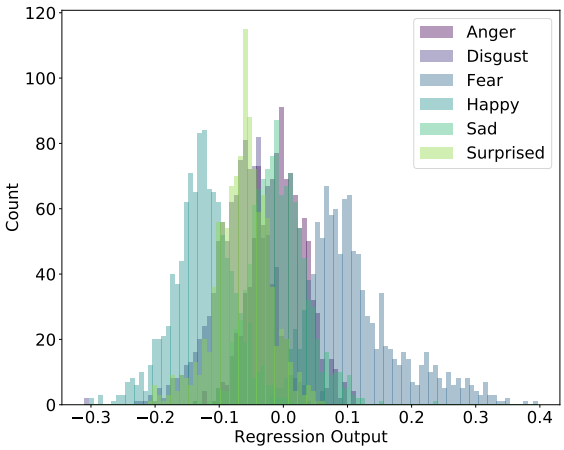

Figure 2: Distribution of predicted intensity targets for each emotion: $[Anger, Disgust, Fear, Happy, Sad, Surprised]$

sity values from the dataset. However, in the presence of relatively high 4-class accuracy, we know that our system is correctly predicting which emotions are present most of the time, and can produce the correct intensity at a coarser-grained step size.

6 Analysis

Unfortunately we were not able to obtain information about the distribution of emotion classes contained in the held-out test set. However, we did observe interesting combinations of emotion label clusters from our training data. More than 70% of the training examples had been labeled with only 1 or 2 emotions, for example: $(Happy, Surprise)$, $(Anger, Disgust)$, $(Disgust, Sad)$ or $(Fear, Sad)$. At the same time, the null case (no emotion) was the second-most prevalent label meaning that many of the video segments in our training data had no emotion at all. There were a few rare cases of interesting combinations, such as all 6 emotions being present in one video segment. This exemplifies the inherent complexity and challenge of human communication and the task of emotion labeling.

In Figure 2, we show the distribution of log-predicted emotion intensities for each of the 6 emotion classes. The BLSTM model appears to have learned a representation where the tuple emotions of $(Surprise, Disgust)$ and $(Anger, Fear)$ each have a similar intensity distribution. Intuitively, this could be justified because these pairs are close to each other on the emotional spectrum, e.g. $Surprise$ is easily mistaken for $Disgust$. Our model however, performs best when distinguishing between $Surprise$ and $Disgust$, implying that although the one-dimensional intensity appears similar in Figure 2, the underlying representation that is learned is complex enough to distinguish between these. At the same time, Figure 2 implies that the model has learned that $Fear$ and $Happy$ are extremely different emotions, seeing as their corresponding distributions are far apart, which is also intuitive.

7 Discussion and Future Work

We have presented our efforts towards creating a robust and effective emotion recognition system. Our best system predicts emotion in video by performing both classification and regression on this challenging multi-label problem. As this is the first grand challenge for this dataset, we were not able to make a direct comparison between other systems at this time. However, our methodology shows that our models improve simply by adding additional modalities. Furthermore, all of our DNN models perform better than chance. To that end, we know that trimodal models perform best, followed by bimodal models and then unimodal models. Our work shows that an early fusion technique can effectively predict the presence of multi-label emotions as well as their coarse-grained intensities. Our approach creates a simple and robust baseline on this new dataset.

In future work, we propose exploring feature selection in order to better understand if and how particular modality features correlate with particular emotions. For example, in the audio modality, a falling pitch might indicate Sad, or a loud volume could indicate $Surprise$. Capturing features that correlate with particular emotions could prove useful for generating emotive speech.

We have shown that this problem benefits from sequence information. Therefore, in future efforts to improve performance, one might explore the distribution of emotions across video segments. It is possible that there are relevant patterns of emotion that are expressed from one segment to the next. A potential approach for this would be to use a fixed-width sliding window across multiple consecutive video segments, and predict emotion labels at regular time intervals.

Acknowledgments

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. The authors would like to thank Steve Renals at University of Edinburgh Centre for Speech Technology Research (CSTR) and the anonymous reviewers for their valuable comments.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A System For Large-Scale Machine Learning. In *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, pages 265–283. USENIX Association.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2014. Continuous Conditional Neural Fields for Structured Regression. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pages 593–608.

Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S Huang. 2016. Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 97–104. ACM.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation (LREC)*, 42(4):335.

Keng-hao Chang, Drew Fisher, and John Canny. 2011. AMMON: A Speech Analysis Library For Analyzing Affect, Stress, and Mental Health on Mobile Phones. *Proceedings of PhoneSense*.

Vladimir Chernykh, Grigoriy Sterling, and Pavel Prihodko. 2017. Emotion Recognition from Speech With Recurrent Neural Networks. *arXiv preprint arXiv:1701.08071*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP a Collaborative Voice Analysis Repository for Speech Technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*, pages 960–964. IEEE.

Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation learning for speech emotion recognition. In *INTERSPEECH*, pages 3603–3607.

Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In *INTERSPEECH 2014, Fifteenth Annual Conference of the International Speech Communication Association (ISCA)*.

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, pages 448–456.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *CoRR*, abs/1408.5882.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Steve Lawrence, C. Lee Giles, Ah Chung Tsoi, and Andrew D. Back. 1997. Face Recognition: A Convolutional Neural-Network Approach. *IEEE Transactions on Neural Networks*, 8(1):98–113.

Jinkyu Lee and Ivan Tashev. 2015. High-Level Feature Representation Using Recurrent Neural Network for Speech Emotion Recognition. In *INTERSPEECH 2015, Sixteenth Annual Conference of the International Speech Communication Association (ISCA)*.

Wootaek Lim, Daeyoung Jang, and Taejin Lee. 2016. Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, pages 1–4. IEEE.

Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian R. Fasel, Mark G. Frank, Javier R. Movellan, and Marian Stewart Bartlett. 2011. The Computer Expression Recognition Toolbox (CERT). In *Ninth IEEE International Conference on Automatic Face and*

Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011, pages 298–305.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Moisés H. R. Pereira, Flávio L.C. Pádua, Adriano C.M. Pereira, Fabrício Benevenuto, and Daniel H. Dalip. November 2016. Fusing Audio, Textual and Visual Features for Sentiment Analysis of News Videos. *Tenth International AAAI Conference on Web and Social Media (ICWSM)*, pages pp. 17–20.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-Level Multimodal Sentiment Analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 973–982.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis. In *Proceedings Empirical Methods in Natural Language Processing (EMNLP)*, pages 2539–2544.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A Survey of Multimodal Sentiment Analysis. *Image and Vision Computing*, 65:3–14.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Samarth Tripathi, Shrinivas Acharya, Ranti Dev Sharma, Sudhanshu Mittal, and Samit Bhattacharya. 2017. Using Deep and Convolutional Neural Networks for Accurate emotion classification on DEAP Dataset. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4746–4752.

Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P. Vlahavas. 2008. Multi-Label Classification of Music into Emotions. In *International Society of Music Information Retrieval (ISMIR)*, volume 8, pages 325–330.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems*, 28(3):46–53.

Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3756–3764.

Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. 2014. Visual Sentiment Prediction with Deep Convolutional Neural Networks. *CoRR*, abs/1411.5731.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *CoRR*, abs/1707.07250.

Amir Zadeh, Paul Pu Liang, Jon Vanbriesen, Soujanya Poria, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. 2018. Multimodal language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Association for Computational Linguistics (ACL)*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *CoRR*, abs/1606.06259.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016b. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6):82–88.

Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.